Projection Pursuit Learning: Approximation Properties

Ying Zhao[1] and Christopher G. Atkeson
MIT Artificial Intelligence Laboratory and the Department of Brain and Cognitive Sciences
NE43-771, 545 Technology Square
Cambridge, MA 02139
617-253-0788
yzhao@bbn.com, cga@ai.mit.edu

Abstract

This paper addresses an important question in machine learning and neural networks: Which neural network architectures work better on which problems? A general approximation method based on a continuous version of projection pursuit regression is developed to show that projection pursuit regression works better on angular smooth functions than on Laplacian smooth functions in high dimensional spaces. Projection pursuit regression can avoid the curse of dimensionality in some cases by learning node functions from data. This has implications for neural network models, as a projection pursuit learning network has a very similar structure to a one hidden layer sigmoidal neural network.

1 Introduction

Projection pursuit is a nonparametric statistical technique to find interesting low dimensional projections of high dimensional data sets. It has been used for nonparametric curve fitting, smoothing, and other data analytic purposes. A projection pursuit learning network (PPLN) is based on projection pursuit regression (PPR) (Friedman and Stuetzle, 1981) and possesses a structure very similar to a one hidden layer sigmoidal neural network (Figure 1). A one hidden layer sigmoidal feedforward neural network approximates a function of a vector $\mathbf{x}$ using the structure:

$$f(\mathbf{x}) = \sum_{j=1}^{n} a_j \sigma(p_j \theta_j^T \mathbf{x} + \delta_j) \tag{1}$$

where σ is a sigmoidal function:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

θ_j are direction parameters with $\|\theta_j\| = 1$, and a_j, p_j and δ_j are function parameters. A

[1]Current address: 70 Fawcett Street, BBN Systems and Technologies, Cambridge MA 02138

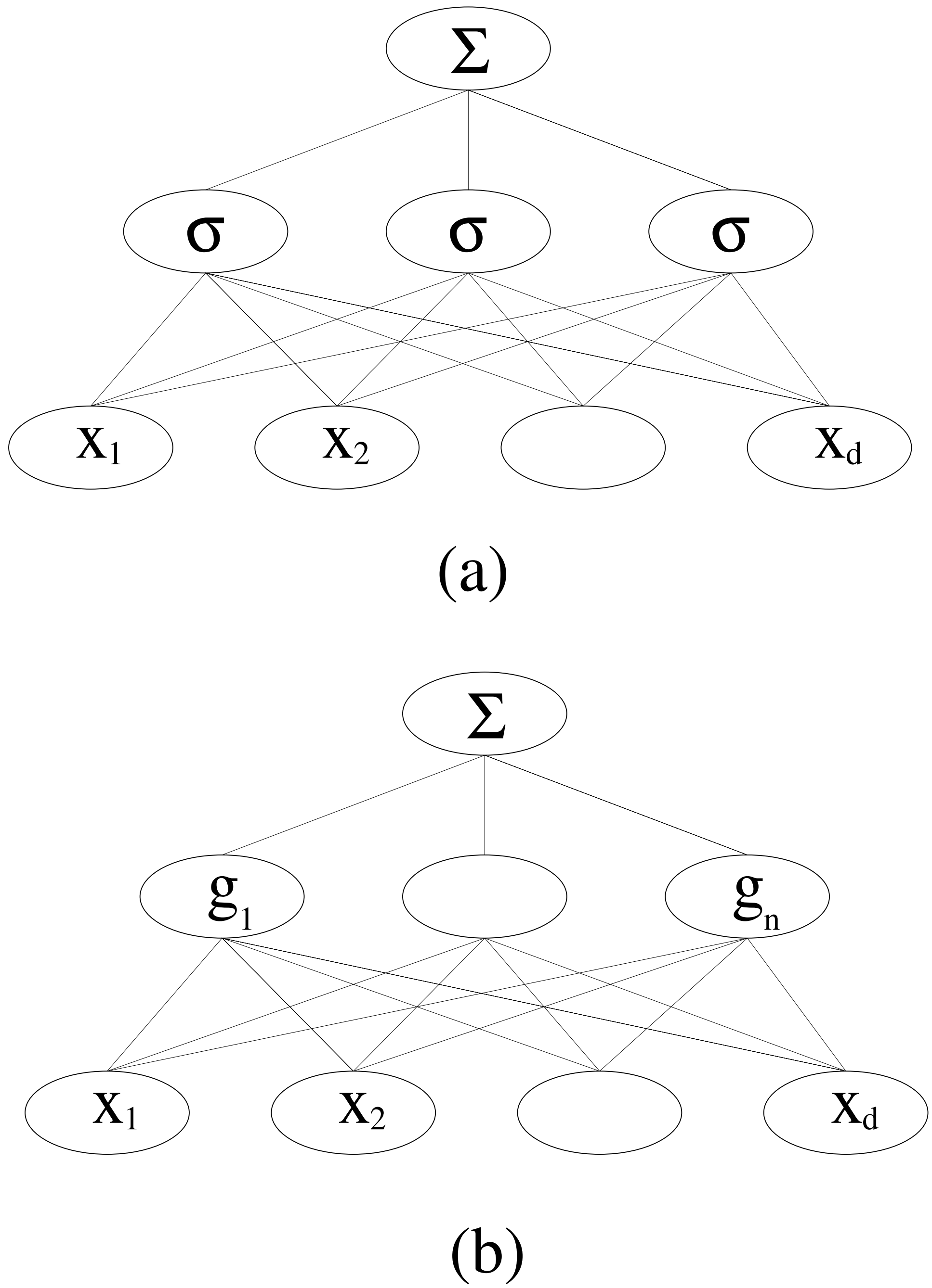


Figure 1: (a) A one hidden layer feedforward neural network. (b) A projection pursuit learning network.

projection pursuit learning network (or ridge function approximation network) approximates a function using the structure

$$f(\mathbf{x}) = \sum_{j=1}^{n} g_j(\theta_j^T \mathbf{x}) \tag{2}$$

where θ_j holds the same meaning as given previously. The corresponding function parameters are implicit in the ridge functions g_j, which are any arbitrary smooth function to be learned from the data. Since the sigmoidal functions are replaced by more general functions g_j, projection pursuit learning networks can be viewed as a generalization of one hidden layer sigmoidal feedforward neural networks. Understanding the properties of a PPLN will help us understand the properties of neural networks. We discuss approximation properties of projection pursuit learning networks in this paper.

Projection pursuit has become one of the most exciting multivariate data analysis methods in statistics since it has the potential to reduce difficulties due to the "curse of dimensionality" in nonparametric statistics by working in low–dimensional linear projections. The original projection pursuit approach, which was called exploratory projection pursuit, was to look for "interesting" low–dimensional projections of higher dimensional data by numerically maximizing a projection index. Projection pursuit techniques were originally proposed by Kruskal (1969, 1972) and related ideas go back to Switzer (1970) and Switzer and Wright (1971). Friedman and Tukey (1974) implemented an early version of projection pursuit. Friedman and Stuetzle extended the ideas behind exploratory projection pursuit and proposed projection pursuit regression (Friedman and Stuetzle, 1981), projection pursuit classification (Friedman 1984b) and projection pursuit density estimation (Friedman, Stuetzle and Schroeder, 1984). Friedman (1984a, 1984c) implemented the supersmoother and backfitting algorithms to perform projection pursuit regression effectively. A review and discussion of projection pursuit was given by Huber (1985). Jones (1987) proved a conjecture about projection pursuit regression in Huber's review paper. Projection pursuit regression was explored theoretically by Diaconis and Shahshahani (1984). They studied necessary and sufficient conditions for functions to be exactly represented as a linear combination of nonlinear ridge functions. Hall (1989) showed explicit formulae for bias and variance of the first projective approximation estimator. Chen (1991) showed that under appropriate conditions the rate of convergence of projection pursuit regression is independent of dimensionality d.

Donoho and Johnstone (1989) demonstrated the duality between projection pursuit regression and kernel regression in two dimensions. Projection pursuit regression behaves well for functions with enough angular smoothness (oscillating slowly with angle), while kernel regression behaves well for functions with enough Laplacian smoothness (oscillations averaging out locally) (more formal definitions are given in Section 3). Thus, projection pursuit learning networks work better on angular smooth functions than on Laplacian smooth functions. Here "work better" means that for fixed complexities of hidden unit functions and a certain approximation accuracy, fewer hidden units are required, or given a fixed number of hidden units, a better accuracy can be achieved. For the two dimensional case ($d = 2$), Donoho and Johnstone (1989) showed this result using equispaced directions. We are interested in generalizing this result to higher dimensions ($d > 2$). There are some difficulties:

- Equispaced directions may not be available when $d > 2$.
- The analysis method in Donoho and Johnstone's paper is limited to two dimensions.

Instead of using equispaced directions, we use uniformly distributed directions on a unit sphere. We apply the theory of spherical harmonics (Muller, 1966) to develop a general method to demonstrate the approximation properties of a PPR scheme which can be applied to any number of dimensions.

Projection pursuit was first introduced in the context of learning networks by Barron and Barron (1988). Barron (1991) showed that the curse of dimensionality can be avoided when sigmoidal neural networks are applied to a special class of underlying functions. We show that the same convergence rate holds when a ridge function approximation scheme is applied to a subset of functions in $L^2(\phi_d)$.

Section 2 shows examples of representing a radial function and a harmonic function by using linear combinations of ridge functions, which are two extreme cases for a PPR scheme. Section 3 proposes a $L^2(\phi_d)$ basis and gives definitions of what is angular and harmonic smoothness. Section 4 examines the case when only one direction is used in approximation and a radial function is the best case to approximate. Section 5 explores a continuous ridge function representation scheme. It is used to explain why a PPR scheme works better on a radial function than a harmonic function when more than one direction is applied in Section 6. Section 7 discusses the curse of dimensionality for a PPR scheme.

2 Examples of Representing Functions by a RA Scheme

Radial functions ($f(\mathbf{x}) = f(|\mathbf{x}|)$) and harmonic functions ($\Delta f = 0$, Δ is the Laplacian operator) are two extreme cases for a projection pursuit regression or a ridge function approximation scheme in higher dimensions. Examples will be shown to illustrate this.

Example 1: a radial function:

$$x^4 + y^4 + z^4 + 2x^2y^2 + 2x^2z^2 + 2y^2z^2 = c_1 \int_{\Omega_3} (\mathbf{x}^T\eta)^4 d\omega_3(\eta) \tag{3}$$

9 directions can represent this three dimensional polynomial exactly. These directions are derived from zeros of Legendre and Tschebyscheff polynomials which are used to discretize the integral on the right side by Gaussian quadrature. Zeros of the Legendre polynomial of degree 3 are:

$$t = 0, \sqrt{\frac{3}{5}}, -\sqrt{\frac{3}{5}}$$

Zeros of the Tschebyscheff polynomial of degree 3 are:

$$s = 0, \frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}$$

The 9 directions are:

$$\begin{bmatrix} \sqrt{1-t_i^2}\sqrt{1-s_j^2} \\ \sqrt{1-t_i^2}s_j \\ t_i \end{bmatrix}$$

where $i = 1,..,3, j = 1,...,3$. The distribution of these directions on a unit sphere is shown in Figure 2 (a).

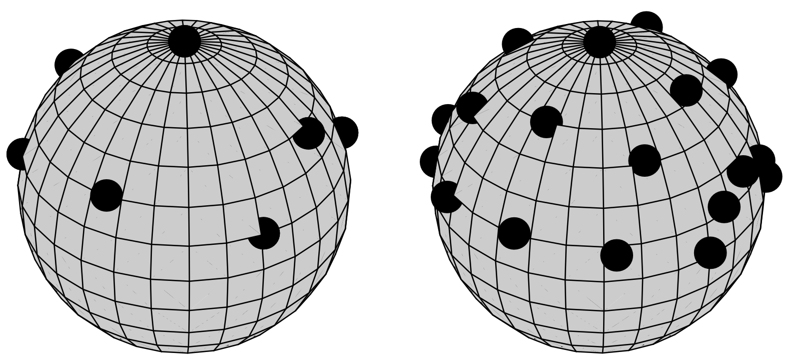
More directions are needed to represent a harmonic function with exactly the same terms as the previous radial function example, but with different coefficients on those terms.

Example 2: a harmonic function

$$\frac{1}{8}(8x^4 + 3y^4 + 3z^4 - 24x^2y^2 - 24x^2z^2 + 6y^2z^2) = c_2 \int_{\Omega_3} (\mathbf{x}^T\eta)^4 S_{4j}(\eta) d\omega_3(\eta) \tag{4}$$

where $S_{4j}(\eta) = \frac{1}{8}(35t^4 - 30t^2 + 3), \eta = t\epsilon_3 + \sqrt{1-t^2}\eta_2$. 25 directions can represent the polynomial, they are:

$$\begin{bmatrix} \sqrt{1-t_i^2}\sqrt{1-s_j^2} \\ \sqrt{1-t_i^2}s_j \\ t_i \end{bmatrix}$$



(a) (b)

Figure 2: Directions (a) for a radial polynomial and (b) for a harmonic polynomial.

where $i = 1, .., 5, j = 1, ..., 5.$ and

$$\begin{aligned} t_i &= 0, 0.90618, -0.90618, 0.53847, -0.53847 \\ s_j &= \cos \frac{2j-1}{10}\pi, j = 1, ..., 5 \end{aligned}$$

The distribution of these directions on a unit sphere is shown in Figure 2(b). More directions are used for the harmonic function than for the radial function.

3 A $L^2(\phi_d)$ Basis and Smoothness Classes

Since smoothness classes typically characterize rates of convergence for approximation schemes, we define some smoothness classes in this section. We first introduce some basic notation. We use $L^2(\phi_d)$ as our underlying function space with Gaussian measure $\phi_d = (\frac{1}{2\pi})^{\frac{d}{2}} e^{-\frac{|\mathbf{x}|^2}{2}}$. Define:

$$\|f\|^2 = \int_{R^d} f^2 \phi_d d\mathbf{x}$$

A function $f(\mathbf{x}) \in L^2(\phi_d)$ means $\|f\|^2$ is bounded. We use polar coordinates here. Let $\mathbf{x} = (x_1, ..., x_d)$ be Cartesian coordinates of a Euclidean space of d dimensions. Then we have

$$|\mathbf{x}|^2 = r^2 = (x_1)^2 + ... + (x_d)^2$$

$$\mathbf{x} = r\xi$$

where

$$\xi = (\xi_1, ..., \xi_d), |\xi| = 1$$

is used to represent a system of coordinates for points on the d dimensional unit sphere Ω_d. The surface element of Ω_d is $d\omega_d$ and the total surface is ω_d, where $\omega_d = \int_{\Omega_d} d\omega_d$.

We derive a complete orthogonal basis in $L^2(\phi_d)$ from the eigenfunctions of a self-adjoint operator. The basis element is defined as:

$$J_{njm}(\mathbf{x}) = \gamma_m \underbrace{r^n S_{nj}(\xi)}_{harmonic} \underbrace{L_m^\alpha(\frac{r^2}{2})}_{radial} \tag{5}$$

where $\mathbf{x} = r\xi$; $n = 0, ..., \infty$; $m = 0, ..., \infty$; $j = 1, ..., N(d,n)$; $\gamma_m = (-2)^m m!$; and $\alpha = n + \frac{d-2}{2}$. $S_{nj}(\xi)$ are linearly independent spherical harmonics of degree n in d dimensions (Muller, 1966). $L_m^\alpha(\frac{r^2}{2})$ is a Laguerre polynomial. The two dimensional version of the basis

J_{njm} appears in Donoho and Johnstone (1989). Deans (1983) discusses some related bases in physics.

Let $\tilde{\Delta}_d$ be a Laplacian operator under standard Gaussian measure ϕ_d. It is a self-adjoint operator.

$$\tilde{\Delta}_d = \frac{1}{r^{d-1}e^{-\frac{r^2}{2}}}\frac{\partial}{\partial r}(r^{d-1}e^{-\frac{r^2}{2}}\frac{\partial}{\partial r}) + \frac{1}{r^2}\Delta_d^* \tag{6}$$

where Δ_d^* is Laplace-Beltrami operator. It can be defined successively with $\Delta_1^* = 0$:

$$\Delta_d^* = (1-t^2)\frac{\partial^2}{\partial t^2} - (d-1)\frac{\partial}{\partial t} + \frac{1}{1-t^2}\Delta_{d-1}^* \tag{7}$$

where $-1 \leq t \leq 1$. The basis J_{njm} can be derived from the eigen function system of $\tilde{\Delta}_d$. Therefore it has the following eigen property:

Lemma 1

$$\tilde{\Delta}_d J_{njm}(r,\xi) = -(n+2m)J_{njm}(r,\xi) \tag{8}$$

The lemma is proved in the appendix.

The advantage of the basis comes from its representation as a product of a spherical harmonic and a radial polynomial. Specifically J_{0jm} is a radial polynomial for $n=0$ and J_{nj0} is a harmonic polynomial for $m=0$. Figure 3(a)-(e) shows these basis elements when $d=2, n+2m=8$.

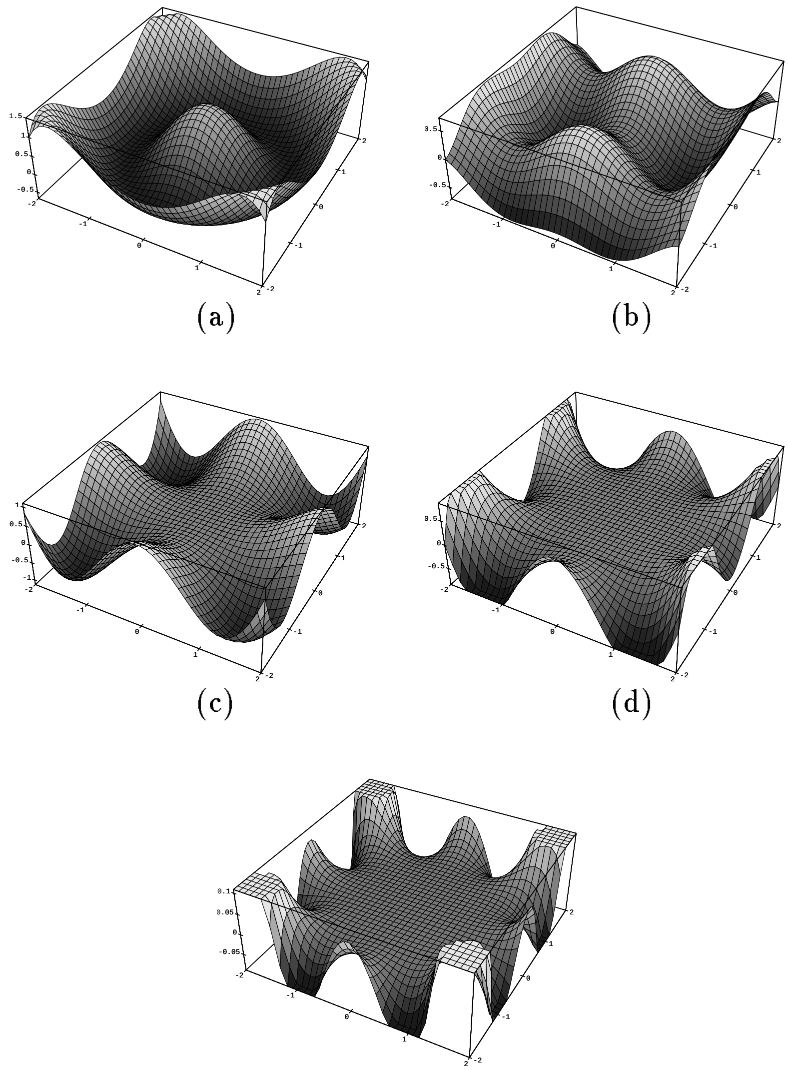
The basis elements J_{njm} have an orthogonality:

$$E(J_{njm}(r,\xi)J_{kil}(r,\xi)) = \delta_n^k \delta_j^i \delta_m^l \frac{2^{n+2m-1}}{\pi^{\frac{d}{2}}}\Gamma(m+n+\frac{d}{2})\Gamma(m+1) \tag{9}$$

where E denotes expectation with respect to ϕ_d. This can be derived from the orthogonalities of Laguerre polynomials and spherical harmonics. Since it is a basis in $L^2(\phi_d)$, any function $f \in L^2(\phi_d)$ has an expansion in terms of basis elements J_{njm}

$$f = \sum_{n,j,m} c_{njm} J_{njm} \tag{10}$$

The circular harmonic $e^{in\theta}$ is a special case of the spherical harmonic $S_{nj}(\xi)$. In two dimensions, $d=2$, $N(d,n)=2$ and $\mathbf{x} = (r\cos\theta, r\sin\theta)$. The spherical harmonic $S_{nj}(\xi)$ can be reduced to the following: $S_{n1}(\xi) = \frac{1}{\sqrt{\pi}}\cos n\theta$, $S_{n2}(\xi) = \frac{1}{\sqrt{\pi}}\sin n\theta$ which is the circular harmonic.



(e)

Figure 3: Basis Elements J_{njm} $d=2, n+2m=8, j=1$ (a) $n=0, m=4$, (b) $n=2, m=3$, (c) $n=4, m=2$, (d) $n=6, m=1$, (e) $n=8, m=0$.

Smoothness classes can be defined in terms of different derivatives:

Definition 1 **Cartesian smoothness of order** *p: if $f(\mathbf{x}) \in L^2(\phi_d)$ has p derivatives and these derivatives are all in $L^2(\phi_d)$.*
Angular Smoothness of order *q: if $\Delta_d^{*q} f \in L^2(\phi_d)$.*
Laplacian Smoothness of order *r: if $\Delta_d^r f \in L^2(\phi_d)$.*

Smoothness classes can also be defined qualitatively from expansions of functions in terms of basis elements J_{njm}. Since $\|f\|^2 = \sum c_{njm}^2 J_{njm}^2$, one can think of $p_{njm}(f) = \frac{c_{njm}^2 J_{njm}^2}{\sum c_{njm}^2 J_{njm}^2}$ as representing the distribution of energy in f among the different modes of oscillation J_{njm}. If f is Cartesian smooth, $p_{njm}(f)$ peaks around small $n + 2m$. If f is angular smooth, $p_{njm}(f)$ peaks around small n. If f is Laplacian smooth, $p_{njm}(f)$ peaks around small m (see Figure 4(a)-(c)).

To explain why PPR works better on angular smooth functions than on Laplacian smooth functions, we first examine how to represent these $L^2(\phi_d)$ basis elements systematically in terms of ridge functions and then use the expansion (10) to derive an error bound of RA for general smooth functions in $L^2(\phi_d)$.

4 Using One Direction for Approximation

It is useful to consider projection pursuit regression using only one direction (hidden unit). Duan (1990) also examined the use of only one projection direction and proposed "adjoint projection pursuit regression" to perform one ridge direction regression. For a $f(\mathbf{x}) : R^d \to R$, the best ridge function approximation to it, given a direction ϵ_d, is the conditional expectation of $f(\mathbf{x})$ in that direction $E(f|\epsilon_d^T \mathbf{x})$ (Huber, 1985). It is a one-variable function defined as:

$$E(f|\epsilon_d^T \mathbf{x}) = \int_{R^{d-1}} f(\mathbf{x}) \rho^*(x_1, ..., x_{d-1}) dx_1 ... dx_{d-1} \tag{11}$$

where $x_i = \epsilon_i^T \mathbf{x}, i = 1, ..., d-1$, $\rho^*(x_1, ..., x_{d-1})$ is the conditional density of a probability density $\rho(\mathbf{x})$.

If only one direction is used to approximate the basis elements J_{njm}, a radial basis is the best one to be approximated by a ridge function when $d = 2$. This is shown in Donoho and Johnstone (1989). It is also the case when $d > 2$. We will see this from the following lemma.

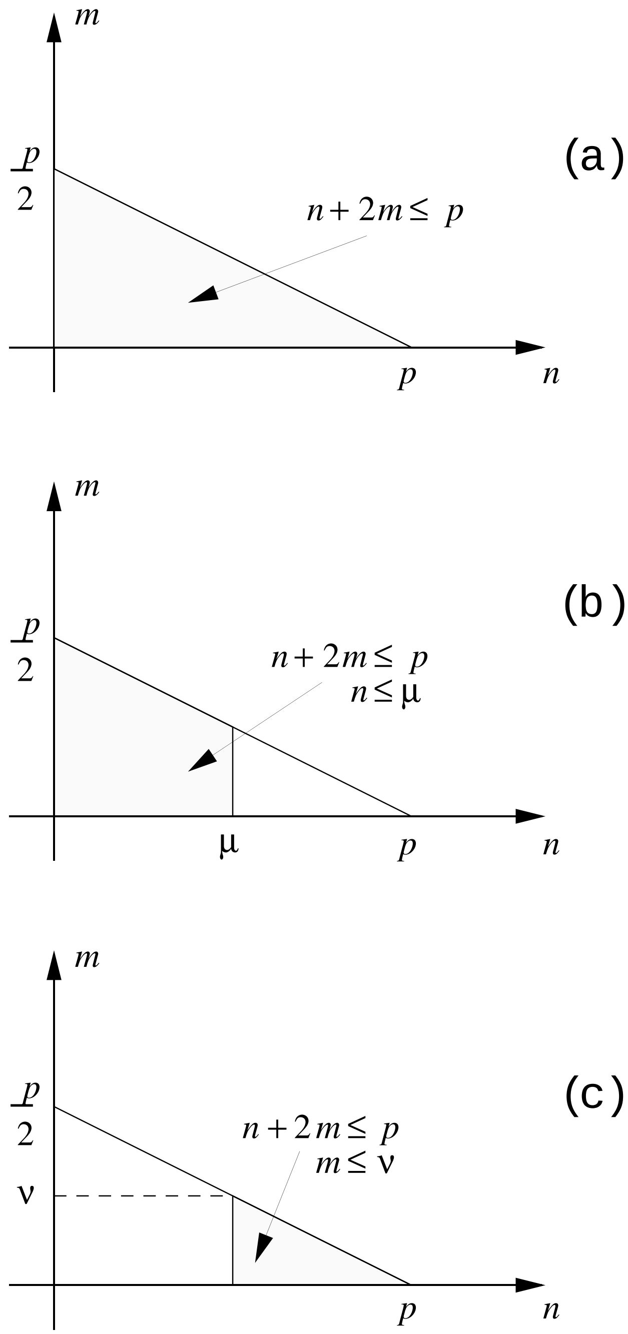


Figure 4: The energy distributions for different smoothness classes. (a) A Cartesian smooth function (b) An angular smooth function (c) A Laplacian smooth function.

Lemma 2 *Let $H_n(x)$ denote Hermite polynomial of degree n. The conditional expectation of the basis element J_{njm} in direction ξ is*

$$E(J_{njm}(\mathbf{x})|\xi^T\mathbf{x}) = S_{nj}(\xi)H_{n+2m}(\xi^T\mathbf{x}) \tag{12}$$

We use the definition for the Hermite polynomials

$$H_m(x) = \frac{1}{\phi(x)}(-D)^m\phi(x) \tag{13}$$

where D denotes differentiation and $\phi(x) = (2\pi)^{1/2}e^{-\frac{x^2}{2}}$. They satisfy the orthogonality relations

$$E[H_k H_l(x)] = k!\delta_l^k \tag{14}$$

In two dimensions $d = 2$, Lemma 2 can be reduced to

$$E(J_{kl}(z,\bar{z})|\xi^T\mathbf{x}) = e^{i(l-k)\theta}H_{l+k}(\xi^T\mathbf{x})$$

which is a fundamental lemma in Donoho and Johnstone (1989).

Suppose the direction η is drawn from the uniform distribution (the density is $1/\omega_d$) on Ω_d. The average relative mean squared error (ARE) can be computed for the first projective approximation $E(J_{njm}|\eta^T\mathbf{x})$ to J_{njm} given the direction η.

$$\begin{aligned}
ARE &= \frac{\int_{\Omega_d}\|J_{njm} - E(J_{njm}|\eta^T\mathbf{x})\|^2\frac{1}{\omega_d}d\omega_d(\eta)}{\|J_{njm}\|^2} \\
&= 1 - \frac{\int_{\Omega_d}\|E(J_{njm}|\eta^T\mathbf{x})\|^2\frac{1}{\omega_d}d\omega_d(\eta)}{\|J_{njm}\|^2} \\
&= 1 - \frac{1}{\lambda_{nmd}\omega_d}
\end{aligned} \tag{15}$$

where the following relation is used:

$$\int_{\omega_d} S_{nj}^2(\eta)d\omega_d(\eta) = 1$$

where

$$\lambda_{nmd} = \frac{2^{n+2m-1}}{\pi^{\frac{d}{2}}}\frac{\Gamma(m+n+\frac{d}{2})\Gamma(m+1)}{\Gamma(n+2m+1)} \tag{16}$$

and

$$\lambda_{nmd}\omega_d = \frac{2^{n+2m}\Gamma(n+m+\frac{d}{2})\Gamma(m+1)}{\Gamma(\frac{d}{2})\Gamma(n+2m+1)} \tag{17}$$

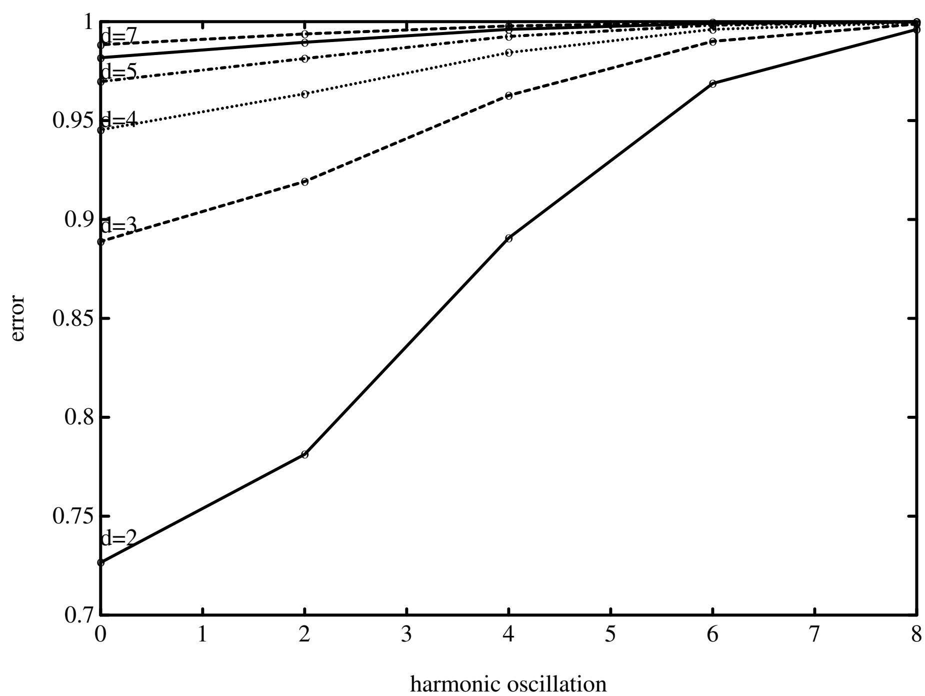


Figure 5: ARE vs. the harmonic oscillation n when only one direction is used.

A plot of ARE *vs* the harmonic oscillation n is shown in Figure 5 for $2 \le d \le 7, n + 2m = 8$. It can be seen from Figure 5 that ARE is minimized when $n = 0$ (a radial element) and maximized when $m = 0$ (a harmonic element) for any $d \ge 2$.

What happens if more than one direction is used? The ridge function approximation scheme still takes advantage of angular smoothness. This can be shown by first considering a continuous ridge function representation and then discretizing it by numerical integration in the following section.

5 The Continuous Ridge Function Representation

There exists a continuous ridge function representation for a smooth function which is an integral of ridge functions through all possible directions.

$$f(\mathbf{x}) = \int_{\Omega_d} g(\mathbf{x}^T\eta, \eta) d\omega_d(\eta). \tag{18}$$

This works intuitively because any object is determined by any infinite set of X-rays taken from all directions. More precisely, any function $f(\mathbf{x})$ with a Fourier transform $h(\omega)$ can be represented as (inverse Fourier transform):

$$\begin{aligned} f(\mathbf{x}) &= \int_{\omega^d} e^{-i\omega\cdot\mathbf{x}} h(\omega) d\omega \\ &= \int_{\Omega_d} \int_{-\infty}^{\infty} e^{-ir(\eta\cdot\mathbf{x})} h(r\eta) r^{d-1} dr\, d\omega_d(\eta) \end{aligned} \tag{19}$$

Therefore $g(\mathbf{x}^T\eta, \eta) = \int_{-\infty}^{\infty} e^{-ir(\eta\cdot\mathbf{x})} h(r\eta) r^{d-1} dr$. The natural discrete approximation to (18) has the form: $f_n(\mathbf{x}) = \sum_{j=1}^{n} w_j g(\mathbf{x}^T\eta_j, \eta_j)$, which becomes the usual PPR. We prove a continuous ridge function representation of basis elements J_{njm} which is shown in Lemma 3.

Lemma 3 *The continuous ridge function representation of J_{njm} is:*

$$J_{njm}(\mathbf{x}) = \lambda_{nmd} \int_{\Omega_d} H_{n+2m}(\eta^T\mathbf{x}) S_{nj}(\eta) d\omega_d(\eta) \tag{20}$$

where λ_{nmd} is in (16).

Therefore any function $f \in L^2(\phi_d)$ has a continuous ridge function representation (18) with

$$g(\mathbf{x}^T\eta, \eta) = \sum c_{njm} \lambda_{nmd} H_{n+2m}(\mathbf{x}^T\eta) S_{nj}(\eta) \tag{21}$$

Different numerical integration schemes can be used to discretize (18) and (20). For example, we use Gaussian quadrature in the examples in Section 2. Since in high dimensions a Monte Carlo integration scheme is the preferred method for numerical integration (Ripley, 1987), we use a Monte Carlo scheme to discretize (18) and (20) in the next section. Uniformly distributed directions used in the Monte Carlo integration are a natural generalization of equispaced directions. If $n + 2m$ is fixed, more function evaluations (or more directions) are needed for a harmonic basis element than for a radial basis element no matter what integration scheme is used.

6 Using Uniformly Distributed Directions on Ω_d for Approximation

Donoho and Johnstone (1989) use equispaced directions in two dimensions. N uniformly distributed directions on Ω_d can be used to generalize equispaced directions which may not exist in higher dimensional spaces. This is a Monte Carlo type of integration scheme on Ω_d.

To approximate the integral (20), N uniformly distributed directions η_1, η_2, ..., η_N on Ω_d drawn independently from the density $f(\eta) = 1/\omega_d$ on Ω_d are used:

$$\hat{J}_{njm}(\mathbf{x}) = \frac{\omega_d}{N}\lambda_{nmd}\sum_{k=1}^{N} H_{n+2m}(\mathbf{x}^T\eta_k)S_{nj}(\eta_k) \tag{22}$$

When the approximation scheme (22) is applied to different basis elements $J_{njm}(\mathbf{x})$, the following result holds:

Theorem 1 *The average relative mean squared error (ARE) for approximating $J_{njm}(\mathbf{x})$ by using N independently drawn uniformly distributed directions on Ω_d and scheme (22) is*

$$ARE = \frac{\int_{\Omega_d} \|J_{njm}(\mathbf{x}) - \hat{J}_{njm}(\mathbf{x})\|^2 \frac{1}{\omega_d} d\omega_d}{\|J_{njm}(\mathbf{x})\|^2} = \frac{\lambda_{nmd}\omega_d - 1}{N} \tag{23}$$

A plot of ARE *vs* the harmonic oscillation n is shown in Figure 6 for $2 \leq d \leq 7, n+2m = 8, N = 2000$. It is easily seen from Figure 6 and (23) that the ARE is minimized when $n = 0$ (a radial element) and maximized when $m = 0$ (a harmonic element) for a fixed $n + 2m = p$. That is, when uniformly distributed directions are used to approximate the basis element J_{njm}, the best case is when the element is radial and the worst case is when the element is harmonic. Comparing Figure 5 and Figure 6, we see that when only one direction is used for approximation, the difference of a harmonic basis and a radial basis element becomes smaller when the dimension d becomes larger, while when uniformly distributed directions are used, the situation is just the opposite.

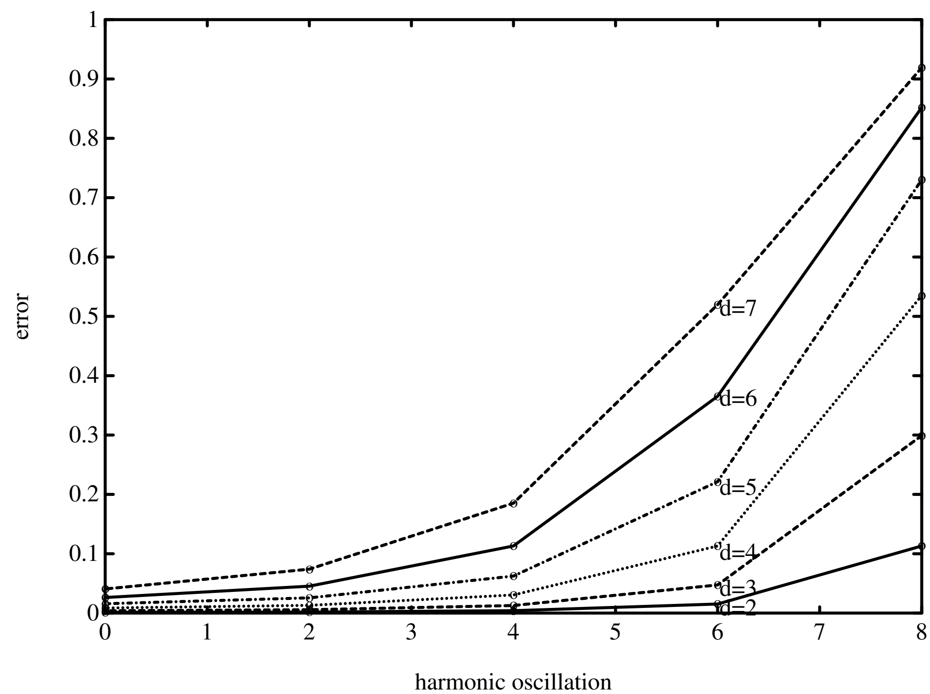


Figure 6: ARE vs. the harmonic oscillation n when only many directions are used.

The same justification can be done for a general function $f \in L^2(\phi_d)$ with (18). A RA scheme is:

$$\hat{f}(\mathbf{x}) = \frac{\omega_d}{N} \sum_{k=1}^{N} g(\mathbf{x}^T \eta_k, \eta_k)$$

$m_N(\mathbf{x}) = f(\mathbf{x})$, $\sigma_N^2 = \frac{\sigma^2}{N}$, and

$$\sigma^2 \quad = \quad \sum c_{njm}^2 \|J_{njm}\|^2 (\lambda_{nmd}\omega_d - 1) \tag{24}$$

Since $\lambda_{nmd}\omega_d$ is small when n is small, large when m is small and recall the distribution $p_{njm}(f)$ in Section 3, $\sigma_N^2/\|f\|^2$ is small when f is smooth. Among these smooth functions, if f

is angular smooth, $\sigma_N^2/\|f\|^2$ is smaller than if f is Laplacian smooth. The error convergence rate $\frac{\sigma_N}{\|f\|} = \frac{\sigma}{\|f\|\sqrt{N}}$ is consequently smaller for f being angular smooth than for f being Laplacian smooth. But both rates are $O(N^{-\frac{1}{2}})$ no matter what class the underlying function belongs to. The difference is the constant which is related to the distribution of energy in f among the different modes of oscillations (angular or harmonic). The radial and harmonic functions are two extremes.

7 The Curse of Dimensionality in RA

Generally, N directions $\eta_1, \eta_2, \ldots\ldots, \eta_N$, on Ω_d drawn independently from the uniform distribution on the sphere Ω_d are used to approximate a function with a continuous ridge function representation (18). A discrete ridge function approximation scheme is:

$$\hat{f}(\mathbf{x}) = \omega_d \frac{1}{N} \sum_{k=1}^{N} g(\mathbf{x}^T \eta_k, \eta_k) \tag{25}$$

The mean value for $\hat{f}(\mathbf{x})$ is $m_N = \omega_d \frac{1}{N} \sum_{k=1}^{N} \int_{\Omega_d} g(\mathbf{x}^T \eta_k, \eta_k) d\omega_d(\eta_k) = f(\mathbf{x})$, and the variance is

$$\sigma_N^2 = \int_{\omega_d} \|\hat{f}(\mathbf{x}) - f(\mathbf{x})\|^2 \frac{1}{\omega_d} d\omega_d(\eta) = \frac{\sigma^2}{N} \tag{26}$$

where

$$\begin{aligned} \sigma^2 &= \int_{\Omega_d} \|\omega_d g(\mathbf{x}^T \eta, \eta) - f(\mathbf{x})\|^2 \frac{1}{\omega_d} d\omega_d(\eta) && (27) \\ &= \omega_d \int_{\Omega_d} \|g(\mathbf{x}^T \eta, \eta)\|^2 d\omega_d(\eta) - \|f\|^2 && (28) \end{aligned}$$

That is, if σ is bounded, $\hat{f}(\mathbf{x}) \to f(\mathbf{x})$ with a rate $O(N^{-\frac{1}{2}})$. Equation (26) shows that there is no curse of dimensionality if a ridge function approximation scheme (25) is used for $f(\mathbf{x})$. The same conclusion can be drawn when sigmoidal neural networks are applied to Barron's class of underlying functions (Barron, 1991). But our function class here is the function class that makes σ in (28) bounded. The following conditions are sufficient for a bounded σ:

1. $f \in L^2(\phi_d)$.

2. f can be represented by a continuous ridge function (18) with $g(\mathbf{x}^T \eta, \eta)$ bounded.

From (19), we know that this function class is a subset of $L^2 \cap L^1$ (excluding those which have unbounded $g(\mathbf{x}^T\eta, \eta)$). In other words, if we are allowed to choose a bounded node function $g(\mathbf{x}^T\eta, \eta)$ on the basis of the data, which is the defining characteristic of PPR, then a ridge function approximation scheme can avoid the curse of dimensionality. This is a generalization of Barron's result that the curse of dimensionality goes away if certain types of node functions (*e.g., cos* and σ) are considered.

8 Conclusion

Since PPR learns bounded node functions from data, there is no curse of dimensionality in terms of an exponentially growing number of hidden units in a projection pursuit learning network. This suggests a projection pursuit network can employ a small number of hidden units but the one dimensional node functions should be parameterized carefully. This corresponds to grouping sigmoidal node functions in a small number of directions in a one hidden layer sigmoidal neural network. In practical algorithms, learning $g(\mathbf{x}^T\eta, \eta)$ is usually replaced by a smoothing step which seeks a one dimensional function to fit $\mathbf{x}^T\eta$ to the residuals in this direction (Friedman and Stuetzle, 1981, Zhao and Atkeson, 1991). The smoothness of an underlying function determines the size of the constant σ^2. As shown in the previous section, if $p(\eta) = 1/\omega_d$ (*i.e.,* uniformly distributed directions), then angular smooth functions have smaller σ^2 than Laplacian smooth functions do. Choosing different $p(\eta)$ does not change this conclusion. But a properly chosen $p(\eta)$ reduces σ^2 in general which corresponds to variance reduction.

9 Acknowledgments

This paper describes research done at the Department of Brain and Cognitive Sciences, the Center for Biological and Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support was provided under Air Force Office of Scientific Research grant AFOSR-89-0500, by the Siemens Corporation, and by the ATR Human Information Processing Research Laboratories. Support for CGA was provided by a National Science Foundation Presidential Young Investigator Award. Special thanks goes to Dr. Zhengfang Zhou at Math Dept. in MIT, who provided useful discussions and reviewed

the draft.

10 Appendix

Proof of Lemma 1 (8):

$$\tilde{\Delta}_d J_{njm}(r,\xi) = \gamma_m S_{nj}(\xi)\left[\frac{1}{r^{d-1}e^{-\frac{r^2}{2}}}\frac{\partial}{\partial r}\left(r^{d-1}e^{-\frac{r^2}{2}}\frac{\partial}{\partial r}r^n L_m^\alpha(\frac{r^2}{2})\right) - n(n+d-2)r^{n-2}L_m^\alpha(\frac{r^2}{2})\right] \tag{29}$$

where

$$\begin{aligned}
& \frac{1}{r^{d-1}e^{-\frac{r^2}{2}}}\frac{\partial}{\partial r}\left(r^{d-1}e^{-\frac{r^2}{2}}\frac{\partial}{\partial r}r^n L_m^\alpha(\frac{r^2}{2})\right) \\
= \; & \frac{1}{r^{d-1}e^{-\frac{r^2}{2}}}\frac{\partial}{\partial r}\left(nr^{n+d-2}e^{-\frac{r^2}{2}}L_m^\alpha(\frac{r^2}{2}) + r^{n+d}e^{-\frac{r^2}{2}}L_m^{\alpha\prime}(\frac{r^2}{2})\right) \\
= \; & -r^{2-d}\left(nr^{n+d-2}L_m^\alpha(\frac{r^2}{2}) + r^{n+d}L_m^{\alpha\prime}(\frac{r^2}{2})\right) \\
& +n(n+d-2)r^{n-2}L_m^\alpha(\frac{r^2}{2}) + (2n+d)r^n L_m^{\alpha\prime}(\frac{r^2}{2}) \\
& +r^{n+2}L_m^{\alpha\prime\prime}(\frac{r^2}{2}) \\
= \; & -nr^n L_m^\alpha(\frac{r^2}{2}) - r^{n+2}L_m^{\alpha\prime}(\frac{r^2}{2}) + n(n+d-2)r^{n-2}L_m^\alpha(\frac{r^2}{2}) \\
& +(2n+d)r^n L_m^{\alpha\prime}(\frac{r^2}{2}) + r^{n+2}L_m^{\alpha\prime\prime}(\frac{r^2}{2}) \\
= \; & -nr^n L_m^\alpha(\frac{r^2}{2}) - r^{n+2}L_m^{\alpha\prime}(\frac{r^2}{2}) + n(n+d-2)r^{n-2}L_m^\alpha(\frac{r^2}{2}) \\
& +2r^n\left(\frac{r^2}{2}L_m^{\alpha\prime}(\frac{r^2}{2}) - mL_m^\alpha(\frac{r^2}{2})\right) \\
= \; & -(n+2m)r^n L_m^\alpha(\frac{r^2}{2}) + n(n+d-2)r^{n-2}L_m^\alpha(\frac{r^2}{2})
\end{aligned}$$

where Laguerre polynomial has the properties:

$$xL_m^{\alpha\prime\prime}(x) + (\alpha+1-x)L_m^{\alpha\prime}(x) + mL_m^\alpha(x) = 0 \tag{30}$$

$$xL_m^{\alpha\prime}(x) = mL_m^\alpha(x) - (m+\alpha)L_{m-1}^\alpha(x) \tag{31}$$

Therefore

$$\tilde{\Delta}_d J_{njm}(r,\xi) = -(n+2m)\gamma_m r^n S_{nj}(\xi)L_m^\alpha(\frac{r^2}{2})$$

Proof of Lemma 2 (12):

Let ξ and η be unit vectors on the sphere Ω_d, *i.e.* ξ, $\eta \in \Omega_d$. The following relations have been used in the proof:

$$E(H_{n+2m}(\eta^T \mathbf{x})|\xi^T \mathbf{x}) = H_{n+2m}(\xi^T \mathbf{x})(\xi^T \eta)^{n+2m} \tag{32}$$

This equation can be derived from D & J's equation (2.4) and in Lemma 2.1.

$$\int_{\Omega_d} (\xi^T \eta)^{n+2m} S_{nj}(\eta) d\omega_d(\eta) = \frac{1}{\lambda_{nmd}} S_{nj}(\xi) \tag{33}$$

where λ_{nmd} is in (16) and equation (33) can be derived from Muller's Lemma 11 & Theorem 6.

$$\begin{aligned}
& E(J_{njm}(\mathbf{x})|\mathbf{x}^T\xi) \\
= \ & E\left(\int_{\Omega_d} \lambda_{nmd} H_{n+2m}(\mathbf{x}^T\eta) S_{nj}(\eta) d\omega_d(\eta) |\mathbf{x}^T \xi\right) && \text{Lemma 2} \\
= \ & \lambda_{nmd} \int_{\Omega_d} E(H_{n+2m}(\mathbf{x}^T\eta)|\mathbf{x}^T\xi) S_{nj}(\eta) d\omega_d(\eta) \\
= \ & \lambda_{nmd} \int_{\Omega_d} H_{n+2m}(\mathbf{x}^T\xi)(\xi^T\eta)^{n+2m} S_{nj}(\eta) d\omega_d(\eta) && \text{from (1)} \\
= \ & \lambda_{nmd} H_{n+2m}(\mathbf{x}^T\xi) \int_{\Omega_d} (\xi^T\eta)^{n+2m} S_{nj}(\eta) d\omega_d(\eta) \\
= \ & H_{n+2m}(\mathbf{x}^T\xi) S_{nj}(\xi) && \text{from (2)}
\end{aligned}$$

Proof of Lemma 3 (20):

Let $\mathbf{x} = r\xi$ and the following equations (34), (35) and (36) have been used in the proof:

$$L_m^\alpha(x) = \sum_{k=0}^{m} (-1)^k \frac{1}{k!} \left(\begin{array}{c} m+\alpha \\ m-k \end{array} \right) x^k \tag{34}$$

$$H_{2m}(\sqrt{x}) = 2^m m! (-1)^m L_m^{-1/2}(\frac{x}{2}) \tag{35}$$

$$\left(\frac{d}{dx}\right)^n H_{n+2m}(x) = \frac{\Gamma(n+2m+1)}{\Gamma(2m+1)} H_{2m}(x) \tag{36}$$

$$
\begin{aligned}
&\int_{\Omega_d} H_{n+2m}(\mathbf{x}^T\eta)S_{nj}(\eta)d\omega_d(\eta) \\
= \;& S_{nj}(\xi)\omega_{d-1}\int_{-1}^{1} H_{n+2m}(rs)p_n(s)(1-s^2)^{\frac{d-3}{2}}d\omega_d(\eta) && \text{Muller, Theorem 6} \\
= \;& S_{nj}(\xi)c\int_{-1}^{1}\left(\frac{d}{ds}\right)^n H_{n+2m}(rs)(1-s^2)^{n+\frac{d-3}{2}}d\omega_d(\eta) && \text{Muller, Theorem 11} \\
= \;& r^n S_{nj}(\xi)c\int_{-1}^{1} H_{n+2m}(rs)(1-s^2)^{n+\frac{d-3}{2}}d\omega_d(\eta) && \text{from (5)} \\
= \;& r^n S_{nj}(\xi)c\int_{-1}^{1} L_m^{-1/2}\left(\frac{r^2s^2}{2}\right)(1-s^2)^{n+\frac{d-3}{2}}d\omega_d(\eta) && \text{from (4)} \\
= \;& r^n S_{nj}(\xi)c\sum_{k=0}^{m}(-1)^k\frac{1}{k!}\left(\begin{array}{c} m-\frac{1}{2} \\ m-k\end{array}\right)\left(\frac{r^2}{2}\right)^k\int_{-1}^{1} s^{2k}(1-s^2)^{n+\frac{d-3}{2}}ds \\
= \;& r^n S_{nj}(\xi)c\sum_{k=0}^{m}(-1)^k\frac{1}{k!}\left(\begin{array}{c} m+\alpha \\ m-k\end{array}\right)\left(\frac{r^2}{2}\right)^k \\
= \;& cr^n S_{nj}(\xi)L_m^{\alpha}\left(\frac{r^2}{2}\right) \\
= \;& cJ_{njm}(x)
\end{aligned}
$$

The constant c can be expressed as

$$
\begin{aligned}
c &= (\frac{1}{2})^n\frac{\Gamma(n+2m+1)}{\Gamma(2m+1)}\frac{\Gamma(n+\frac{d-1}{2})}{\Gamma(m+\alpha+1)}\omega_{d-1} \\
&= \frac{\pi^{\frac{d}{2}}\Gamma(n+2m+1)}{2^{n+2m-1}\Gamma(m+1)\Gamma(m+\alpha+1)} \\
&= \frac{1}{\lambda_{nmd}}
\end{aligned}
$$

Proof of Theorem 1:

Samples $\eta_i, i=1,...,N$ are *i.i.d* random variables drawn from the density $1/\omega_d$ on Ω_d and form (22). The mean value for $\hat{J}_{njm}(\mathbf{x})$ is

$$
m_N(\mathbf{x}) = \frac{\omega_d}{N}\lambda_{nmd}\sum_{k=1}^{N}\int_{\Omega_d} H_{n+2m}(\mathbf{x}^T\eta_k)S_{nj}(\eta_k)\frac{1}{\omega_d}d\omega_d(\eta_k) = J_{njm}(\mathbf{x}) \tag{37}
$$

The variance is

$$
\sigma_N^2 = \int_{\Omega_d}\|J_{njm}(\mathbf{x})-\hat{J}_{njm}(\mathbf{x})\|^2\frac{1}{\omega_d}d\omega_d = \frac{\sigma^2}{N} \tag{38}
$$

where

$$\begin{aligned}
\sigma^2 &= \int_{\Omega_d} \|\lambda_{nmd}\omega_d H_{n+2m}(\mathbf{x}^T\eta)S_{nj}(\eta) - J_{njm}(\mathbf{x})\|^2 \frac{1}{\omega_d} d\omega_d(\eta) \\
&= \frac{1}{\sqrt{(2\pi)^d}} \int_{R^d} e^{-\frac{r^2}{2}} \int_{\Omega_d} \\
&\quad \left[\lambda_{nmd}\omega_d H_{n+2m}(\mathbf{x}^T\eta)S_{nj}(\eta) - J_{njm}(\mathbf{x})\right]^2 \frac{1}{\omega_d} d\omega_d(\eta) d\mathbf{x} \\
&= \frac{1}{\sqrt{(2\pi)^d}} \int_{R^d} e^{-\frac{r^2}{2}} \\
&\quad \left[\int_{\Omega_d} \lambda^2_{nmd}\omega_d H^2_{n+2m}(\mathbf{x}^T\eta)S^2_{nj}(\eta) d\omega_d(\eta) - J^2_{njm}(\mathbf{x})\right] d\mathbf{x} \\
&= \lambda^2_{nmd}\omega_d \int_{\Omega_d} E(H^2_{n+2m}(\mathbf{x}^T\eta))S^2_{nj}(\eta) d\omega_d(\eta) - \|J_{njm}\|^2 \\
&= \lambda^2_{nmd}\omega_d \Gamma(n+2m+1) - \|J_{njm}\|^2 \\
&= \frac{\|J_{njm}\|^4 \omega_d}{\Gamma(n+2m+1)} - \|J_{njm}\|^2 \\
&= \|J_{njm}\|^2 \left[\frac{\|J_{njm}\|^2 \omega_d}{\Gamma(n+2m+1)} - 1\right] \\
&= \|J_{njm}\|^2 (\lambda_{nmd}\omega_d - 1) \qquad (39)
\end{aligned}$$

Therefore,

$$ARE = \frac{\sigma_N^2}{\|J_{njm}\|^2} = \frac{\lambda_{nmd}\omega_d - 1}{N}$$

11 References

Barron, A. R. and Barron, R. L. (1988) "Statistical Learning Networks: AUnifying View". in *Computing Science and Statistics: Proceedings of 20th Symposium on the Interface.* Wegman, editor, *American Statistical Association*, Washington, D.C., 192-203.

Barron, A. R. (1991) "Universal Approximation Bounds for Superpositions of a Sigmoidal Function". Technical Report No. 58. Department of Statistics, University of Illinois at Urbana-Champaign.

Chen, H. (1991) "Estimation of a Projection-pursuit Type Regression Model". *The Annals of Statistics*, 19: 142-157.

Deans, S. R. (1983), *The Radon Transform and Some of Its Applications.* Wiley, New

York.

Diaconis, P. and Shahshahani, M. (1984) "On Non-linear Functions of Linear Combinations". *SIAM Journal of Scientific and Statistical Computing*, 5, 175-191.

Donoho, D. L. and Johnstone, I. (1989) "Projection-based Approximation, and Duality with Kernel Methods." em The Annals of Statistics, 17, 58-106.

Duan, N. (1990) "The Adjoint Projection Pursuit Regression." *Journal of the American Statistical Association*, Vol. 85, No. 412, pp 1029-1038.

Friedman, J. H. and Stuetzle, W. (1981) "Projection Pursuit Regression". *Journal of American Statistical Association*, 76: 817-823.

Friedman, J. H. and Tukey, J. W. (1974) "A Projection Pursuit Algorithm for Exploratory Data Analysis". *IEEE Transactions on Computers*, 23: 881-889.

Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984) "Projection Pursuit Density Estimation". *Journal of the American Statistical Association*, 79: 599-608.

Friedman, J. H. (1984a) "A Variable Span Smoother". Department of Statistics. Stanford University, Report LCM005.

Friedman, J. H. (1984c) "Smart User's Guide". Department of Statistics,Stanford University, Report LCM001.

Friedman, J.H. (1984b) "Classification and Multiple Response Regression Through Projection Pursuit". Dept. of Statist., Stanford University,Report LCM006.

Hall, P. G. (1989) "On Projection Pursuit Regression". *The Annals of Statistics*, 17(2): 573-588.

Huber, P. J. (1985) "Projection Pursuit (with discussion)." *The Annals of Statistics*, 13, 435-475

Jones, L. (1987) "On a Conjecture of Huber Concerning the Convergence of Projection Pursuit Regression." *The Annals of Statistics*, 15, 880-882.

Kruskal, J. B. (1969) "Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation Which Optimizes a New Index of Condensation'." in *Statistical Computation.* (R. C. Milton and J. A. Nelder, eds.) Academic Press, New York.

Kruskal, J. B. (1972) "Linear Transformation of Multivariate Data to Reveal Clustering." in *Multidimensional Scaling : Theory and Applications in the Behavioral Sciences, I, Theory.* Seminar Press, New York and London.

Muller, C. (1966) *Spherical Harmonics.* Lecture Notes in Mathematics, no.17.

Ripley, B. (1987) *Stochastic Simulations*, John Wiley & Sons, New York.

Switzer, P. and Wright, R.M. (1971) "Numerical Classification Applied toCertain Jamaica Eocene Nummulitids". *Mathematical Geology*, 3, 297-311.

Switzer, P. (1970) "Numerical Classification". in *Geostatistics.* Plenum, New York.

Zhao, Y., Atkeson, C. G. (1991) "Projection Pursuit Learning", in *Proc. IJCNN-91-Seattle.* A more extensive manuscript is in preparation.